事件相关电位研究的统计检验力分析：方法及影响因素

念靖晴 [1*] 陈曦 [2] 陈芳芳 [3] 牛霞 [4] 罗禹 [1*]

([1]贵州师范大学心理学院，贵阳，550025; [2]上海 OPPO 公司，上海，200032;[3]芜湖市第四人民医院，芜湖，241003; [4]安徽医科大学，合肥，230031)

摘 要 研究结果的稳健性和可重复性对于科学研究的发展至关重要，但在事件相关电位研究文献却极少见到完整的统计检验力报告。本文主要是通过对已有研究的梳理总结，从而介绍事件相关电位研究中统计检验力分析方法、应用实例以及实验设计、效应幅值、样本量以及试次数量等影响因素，以期为研究者设计和/或预注册研究方案等阶段计算和报告事件相关电位研究中的统计检验力提供参考。

关键词 脑电 事件相关电位 统计检验力 样本量 试次数量

1 引言

在心理学研究可重复性危机背景下(聂丹丹 等, 2016; 胡传鹏 等, 2016)，研究结果的稳健性（robust）和可重复性(reproducibility)对于心理学研究的发展至关重要。研究发现，统计检验力（statistical power）决定了研究结果的置信水平，是衡量其研究结果可靠性和研究可重复性的关键指标之一(Fraley & Vazire, 2014; Schweizer & Furley, 2016)，在研究结果的稳健性和可重复性中起着决定性作用。统计检验力是指当零假设（null hypothesis）为假时，统计测验正确拒绝零假设的概率，一般用 1-β 表示，通常设置为 0.8(Cohen, 1988, 2013)。与统计检验力高的研究相比，统计检验力低的研究会导致更多的假阳性（Type-I error，I 类错误）和假阴性（Type-II error，II 类错误）结果。然而在过去的 60 年中，科学研究领域的统计检验力约为 24% (Smaldino & McElreath, 2016)。其中，神经科学研究领域的统计检验力在 8%~30% 范围之间(Button et al., 2013)，远远低于统计检验力要达到至少 80% 的理想水平。因此，研究者越来越担心由于统计检验力不足，可能导致了大多数科学研究的结果是虚假的(Ioannidis, 2005; Munafò et al., 2017)。

脑电技术是认知神经科学领域中极为重要和被研究者广泛使用的研究工具之一。但脑电研究文献却极少见到完整的统计检验力报告，Clayson 等人（2019）研究发现仅仅只有 15% 的脑电研究得到了合适的统计检验力(Clayson et al., 2019)。一方面可能与认知神经科学领域（特别是电生理技术领域）传统以来的常用研究范式有关。目前，心理学的很多研究研究（如：事件相关电位研究）范式提倡反复测量被试在特定条件下的反应，即：对同一刺激类型的反应进行多个试次的测量，随后对多次测量结果进行平均，以期达到对被试真实反应地更精确估计。另一方面也可能与脑电研究复杂的数据结构有关，例如，原始的单通道

*通讯作者：念靖晴，E-mail:nianjingqing@126.com; 罗禹，E-mail: yuluo@gznu.edu.cn

脑电信号是三维数据，具有频率（Frequency）、时间（Time）和电压振幅（Amplitude）三个维度；研究者可能会对在长频率、特定时间上脑电数据感兴趣（时域分析）、也可能对特定频率，长时间段的脑电数据感兴趣（频域分析）、或是对两者同样感兴趣（时频分析）等，从而导致传统的统计检验力分析方法难以准确适用。在本文中，我们将仅探讨时域分析中典型的研究方法：事件相关电位（Event-related potential, ERP）。

研究发现，ERP 研究中的统计检验力会受到实验设计（study design）、效应幅值（effect magnitude）、样本量（sample size）以及试次数量（number of trials）等因素的影响(Boudewyn et al., 2018; Gibney et al., 2020; Hall et al., 2023; Jensen & MacDonald, 2023; Ngiam et al., 2021)。其中，实验设计特指实施实验处理的一个计划方案以及与计划方案有关的统计分析；效应幅值是指以微伏为单位效应的绝对值大小；样本量是指参与研究的人员数量；试次数量是指研究者能够采集到符合研究需求数据的相对较少试次数。

样本量和试次数量在脑电实验设计时起着重要的作用。Clayson 等人（2019）指出在事件相关电位研究中通过对样本量和试次数量进行先验分析，可以在一定程度上确保适宜的统计检验力和实验结果的稳健性，从而降低研究的可重复性危机。然而已有的研究较多关注样本量对统计检验力的影响，而忽略了试次数量的影响。同时已有的大多数研究者往往使用经验法则而非遵循固定标准来确定研究中的试次数量。因此，研究者应当遵循什么样的标准来决定需要多少名被试（样本量），以及每个被试完成多少个试次（试次数量）目前仍不明确。试次数量可以类比成调查科学中的量表题目数（基于当下目前仍然流行的研究范式经典测量理论），基于结构方程模型和心理测量学的大量研究或范式已经倡议在研究中要使用足够数量的题项，从而提高测量的统计检验力(McQuitty, 2004; Zhang & Stone, 2008)。然而，事件相关电位研究则没有一个明确的公式或计算方法，仅有模糊的，且不同研究组织/团体有着不同的经验标准。同时，也较少有研究者关注事件相关电位研究中实验设计和效应幅值对统计检验力的影响。比如：在进行被试内或被试间实验设计时，确定能稳健地分离出每个实验处理水平之间反应差异的样本量，以及实验处理水平之间效应幅值稳定可信的试次数量。此外，研究人员经常需要在样本量和试次数量之间进行权衡。具体来说，就是由于时间、科研经费等客观因素的影响，研究人员经常需要在增加样本量减少试次数量或者减少样本量增加试次数量之间进行选择。然而这种权衡的决策标准以及这种权衡对统计检验力的影响尚不明确。因此，如果不系统地研究样本量、试次数、效应幅值以及实验设计等因素如何影响事件相关电位研究中的统计检验力，就很难得出稳健可信的结论。

值得注意的是，事件相关电位研究领域的研究者似乎没有完全适应或认同在研究中需要完整报告统计检验力的这种做法，目前仍有很多研究者在进行事件相关电位研究时并没有提及统计检验力的有关内容。Larson 和 Carbine（2017）发现大多数事件相关电位研究没有报告统计检验力的先验计算指标或者需要其他人自行计算所需的信息。近年来，研究者

开始以事件相关电位研究中试次数量确定依据为切入点，通过模拟数据的方式，系统地探讨实验设计、效应幅值、样本量以及试次数量等因素对统计检验力的影响(Boudewyn et al., 2018; Gibney et al., 2020; Hall et al., 2023; Jensen & MacDonald, 2023; Ngiam et al., 2021)。本文主要是通过对已有研究的梳理总结，从而介绍事件相关电位研究中统计检验力分析方法、应用实例以及影响因素等。

2 事件相关电位研究中统计检验力分析方法及应用实例

2.1 数据驱动法（data-driven method）

数据驱动法的目的是确定在事件相关电位研究中获得特定稳健 ERP 成分所需的最少试次数量。该方法的具体步骤是：将已经获得稳健 ERP 成分的试次数量作为总体，随后从总体中抽取一定数量的试次作为样本，随后对样本进行平均，并将平均样本数据后 ERP 成分与总体样本的 ERP 成分进行对比。不断重复上述过程，直到确定在样本中得到与总体样本相当的 ERP 成分，并确定样本的试次数量，该试次数量大小即为获得该 ERP 成分所需的最少试次数量。总体 ERP 成分与样本 ERP 成分的相似性通过相关系数、内部一致性系数(Olvet & Hajcak, 2009; Thigpen et al., 2017)、重测信度(Huffmeijer et al., 2014; Segalowitz & Barnes, 1993)以及等值性(Marco-Pallares et al., 2011; Pontifex et al., 2010)等指标进行评估。数据驱动法能确定一个稳健的 ERP 成分以及获得它所需的最低试次数量，但其不适用于确定在不同实验处理水平或被试间等的 ERP 成分是否有差异时的试次数量。

在应用实例方面，数据驱动法被运用于 ERP 研究领域中确定 error-related negativity (ERN), error positivity (Pe), N100, N200, vertex positive potential (VPP)/N170, mismatch negativity (MMN), feedback-related negativity (FRN), late positive potential (LPP), and P300 等 ERP 成分的试次数量(Cohen & Polich, 1997; Duncan et al., 2009; Fischer et al., 2017; Huffmeijer et al., 2014; Larson et al., 2010; Marco-Pallares et al., 2011; Olvet & Hajcak, 2009; Pontifex et al., 2010; Rietdijk et al., 2014; Segalowitz & Barnes, 1993; Steele et al., 2016; Thigpen et al., 2017)。

2.2 蒙特卡洛模拟（Monte Carlo analyses）

蒙特卡洛模拟的主要原理是通过指定虚拟总体（分布）以生成虚拟样本（抽样）。具体而言，通过从虚拟总体（分布）中对被试数量和试次数量进行重抽样，从而模拟具有不同试次数量、样本量、效应幅值以及实验设计的实验。在事件相关电位研究的蒙特卡洛模拟中，研究者使用采集到的真实脑电数据作为指定总体。并添加了人工效应（artificial effects），从而为被试内和被试间的分析获取真实的效应幅值(Kiesel et al., 2008; Smulders, 2010; Ulrich & Miller, 2001)。相比于数据驱动法，蒙特卡洛模拟较为理想，因为其结合了真实的脑电数据（噪音的真实性）和人工诱发的实验效应（结果的真实）。通过对每个给定参数集模拟 1000 次实验，研究者能够估计每个参数组合在 α=0.05 水平下获得显著统计结果的概率（即：统计检验力）。随后，研究者使用 *t* 检验确定既定参数组合模拟生成的 ERP

成分是否在条件之间（配对样本 *t* 检验）或组别之间（独立样本 *t* 检验）是否有显著差异。

在应用实例上，蒙特卡洛模拟分析被运用于事件相关电位研究领域中 LRP、ERN、N170、 MMN、P3、N2pc、N400、CDA、N1、Tb、P2 等 ERP 成分的统计检验力分析 (Boudewyn et al., 2018; Gibney et al., 2020; Hall et al., 2023; Jensen & MacDonald, 2023; Ngiam et al., 2021)。同时，为了能让研究者在实际研究中应用该方法，Hall 等人（2023）提供了在线程序 Erp Power Calculator（访问链接为：https://bradleynjack.shinyapps.io/ErpPowerCalculator/），事件相关电位研究中听觉领域的研究者可以通过选择特定的 ERP 成分（N1/Tb/P2）、试次数量（20~1000）、样本量（10~100）、效应幅值（0~3 μV）、实验设计（被试内/被试间）、alpha 水平（0.05/0.01/0.005/0.001）等参数来计算研究的统计检验力。在视觉工作记忆领域，Ngiam 等人（2021）提供了在线程序 CDA Power Calculator（访问链接为：https://williamngiam.shinyapps.io/CDAPower/），可以通过选择感兴趣的效应（稳健 CDA 成分/记忆负荷 2 v.s 4/记忆负荷 2 v.s 6），灵活调整样本量、干净试次数量、统计检验力等参数之间的组合来计算相应的指标。Jensen 和 MacDonald（2023）在 OSF 平台（访问链接为：https://osf.io/wv3da/）公开共享了基于 ERP CORE 数据集对 LRP、ERN、N170、 MMN、P3、N2pc、N400 七个 ERP 成分通过动态组合样本量、试次数量、效应幅值以及实验设计等参数模拟计算统计检验力的代码资源。

2.3 功效等值线图（Power Contours Plot）

功效等值线图是统计检验力的二维平面表征，是样本量（N）和试次数量（k）的联合函数，并可以在其它约束条件下进行优化(Baker et al., 2021)。其核心步骤是，在给定的统计检验力条件下，充分考虑方差的影响，动态调整样本量和试次数量并计算相应的统计检验力，直到计算的结果值符合预设值。使用功效等值线图可以在样本量和试次数量的权衡过程中找到一个决策边界的功率等值线拐点，从而根据实际情况选取适宜的样本量和试次数量。但该方法也有一定局限性，因为在事件相关电位研究中统计检验力不仅仅取决于样本量和试次数量，也取决于研究中特定脑电成分在条件间的效应幅值(Boudewyn et al., 2018)。

在应用实例上，功效等值线图被运用于事件相关电位研究领域中 P100、P200、N600 等 ERP 成分的样本量和试次数量的决策拐点计算(Baker et al., 2021)。同时，为了方便研究者使用该方法来确定实际研究中的样本量和试次数量，Baker 等人（2021）等人开发了在线程序 Power contour estimation（访问链接为：https://shiny.york.ac.uk/powercontours/），通过输入样本量、试次数量、alpha 水平、均值差异、被试内标准差、被试间标准差、招募成本等参数来计算研究的统计检验力，以及实际研究中样本量和试次数量权衡的决策拐点。

3 事件相关电位研究中统计检验力分析的影响因素

在事件相关电位研究中，叠加平均（如：试次数量等）是得到 ERP 成分常用的分析方法之一。在进行统计检验时，GLM（ANOVA）是常用的统计分析方法。因此样本量、试

次数量、效应幅值以及实验设计等因素会影响事件相关电位研究中的统计检验力。

3.1 样本量

样本量是统计检验力函数的直接参数，其增加会显著提高研究统计检验力。在 ERP 研究中，小样本量无疑是损害统计检验力的直接要素。例如：Gibney 等人（2020）研究发现，在被试间实验设计中，若每组样本量为 10 人，则产生真实显著结果的可能性极低。总的来说，与增加试次数量相比，增加样本量对统计能力的影响要更大。此外，尽管试次数量也是影响统计检验力的重要因素，增加试次数量可能也会对提升统计检验力有所帮助，但在实际情况允许的条件下，优先考虑样本量的增加可能会是更优的选择。

3.2 试次数量

试次数量对统计检验力的影响取决于效应量（effect size）的变化。随着试次数量的增加，试次间变异性降低，从而导致效应量增加和统计检验力提升。研究显示，在样本量不充足且检测效应量中等的情况下，试次数量提高约一倍左右能有效地提升统计检验力，使其达到合适的水平 (Boudewyn et al., 2018)。

3.3 效应幅值

先前研究发现，效应幅值较大的 ERP 成分往往需要的试次数量会较少(Baker et al., 2021; Boudewyn et al., 2018)。例如：若被试内实验设计中条件之间的效应幅值很大时，样本量和试次数量的变化对统计检验力的影响较小；当效应幅值在中等水平时，样本量和试次数量的变化对统计检验力的变化有很大的影响；此外，若试次数量足够大，在效应幅值较小时，也能够通过增加样本量以达到足够的统计检验力。

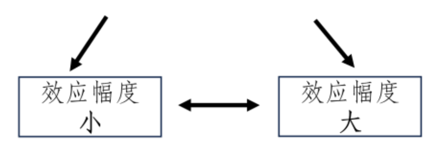
3.4 实验设计

研究发现，与被试间实验设计相比，在被试内实验设计中，试次数量的变化会对统计检验的影响更大，并且在只需要较少的样本量和试次数量就能获得在特定效应大小与被试间实验设计相同的统计检验力水平。而被试间实验中，样本量的变化对统计检验力的影响更明显。先前研究发现，在多数情况下，在被试内实验设计的数据模拟中，将试次数量加倍可以将统计检验力提升至少 1 倍，而样本量加倍的影响则不明显。相比之下，在被试间实验设计的数据模拟中，样本量加倍比试次数量加倍对统计检验力的影响更明显。

被试内

优先考虑增加试次数量，在条件允许的情况下，也可以增加样本量。

*在P3成分上，增加试次数量和样本量的考量应该是一致的。



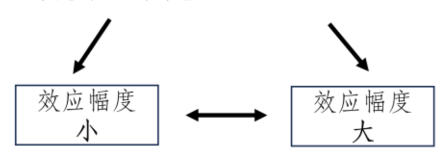
优先增加试次数量的作用比增加样本量大。

增加样本量获得差异显著结果的的可能性更大，即使试次数量没有变化。

被试间

优先考虑增加样本量，在条件允许的情况下，也可以增加试次数量。

*在N170和N2pc成分上，应该优先考虑增加样本量。



优先增加样本量的作用比增加试次数量大。

增加试次数量获得差异显著结果的的可能性更大，即使样本量没有变化。

图 1 被试内实验设计和被试间实验设计模拟结果的部分关键内容示意图。示意图描述了现有研究中统计检验力影响因素之间的动态关系，可以为研究者提供基本指导，帮助研究者了解研究设计中应该关注什么，从而得到更高的统计检验力。请注意，该图描述了观察到的具有更多中等效应幅值的关系。地板效应表明，如果效应幅值太小，那么增加样本量或试次数量没有太大作用。同样，天花板效应表明，样本量和试验次数的增加可能也不会对非常大的效应幅值产生太大影响（尽管这种规模的效应幅值在实际研究中可能并不存在）。改编自：(Jensen & MacDonald, 2023)

4 事件相关电位研究中统计检验力分析的挑战

研究者通过模拟数据的方式，系统地探讨实验设计、效应幅值、样本量以及试次数量等因素通过交互方式对统计检验力产生影响。但在未来的研究中还应该关注：

关注研究中可能出现的天花板效应（ceiling effect）和地板效应（floor effects）。先前研究发现，统计检验力会随着样本量和试次数量的变化而变化，而当统计检验力出现天花板效应或地板效应时，样本量和试次数量的变化并不会对统计检验力的影响就微乎其微了。

此外，关注事件相关电位研究中信噪比（Signal-Noise Ratio, SNR）对统计检验力的影响。信噪比是指脑电数据中信号水平与噪声水平的比值。在脑电数据中，若噪声水平增加，则信噪比和统计检验力都会降低。然而，已有的数据模拟中，无法有效的实际模拟出每个脑电数据中真实的信噪比水平。此外，事件相关电位研究中的信噪比会受到脑电数据采集（如：不同的采集环境和设备、电阻水平等）(Kappenman & Luck, 2010; Laszlo et al., 2014; Luck & Kappenman, 2017; Picton, 2010; Puce & Hämäläinen, 2017)、处理方法(Clayson et al., 2021; Delorme, 2023; Sandre et al., 2020)以及统计检验方法(Luck & Gaspelin, 2017)的影响。Luck 和 Gaspelin（2017）研究发现事件相关电位研究中数据分析的研究者自由度（如：不同的处理与分析管道等）可能会导致假阳性结果。因此，信噪比对事件相关电位研究中统计功效的影响是未来研究探索的一个重要方向。

同时，需要在更复杂的实验情境进一步验证。现有的研究模拟了被试内和被试间实验

设计中试次数量、样本量以及效应幅值是如何影响统计检验力的，但这些已有的结论是否适用于更复杂的实验设计（如：混合实验设计等）、分析方法（如：多因素分析、大规模单变量分析、混合线性模型等）以及不同的样本群体、实验刺激或实验范式中，仍需进一步验证。

在进行结果推广时应持一种审慎态度。因为目前现有的结果都是在一种模拟的理想状态，因此它们可能无法推广到与模拟计算数据集显著不同的情况。同时，虽然有一些研究者开发并提供了统计检验力的计算工具，但其计算工具都是针对事件相关电位研究中特定的 ERP 成分，在未来的研究中应该开发一个更具有广泛适用性的统计检验力计算工具。

5 总结

为了提升事件相关电位研究领域实验结果的稳健性和可重复性，研究者在设计和/或预注册研究方案阶段，需要着重考量统计检验力及实验设计、效应幅值、样本量以及试次数量等因素的影响，从而不断优化研究方案，减少投入在统计检验力不足研究上成本的可能性；同时也能鼓励研究人员神经科学研究报告完整的统计检验力，不断提高科学研究的严谨性和可重复性。值得注意的是，当前结论来源于数据模拟的理想状态，研究人员参考时需结合实际的研究情境。

参考文献

聂丹丹, 王浩, 罗蓉. (2016). 可重复性:心理学研究不可忽视的实践. 中国临床心理学杂志, *24*(4), 618–622. https://doi.org/10/gspvwn

胡传鹏, 王非, 过继成思, 宋梦迪, 隋洁, 彭凯平. (2016). 心理学研究中的可重复性问题:从危机到契机. 心理科学进展, *24*(9), 1504–1518.

Baker, D. H., Vilidaite, G., Lygo, F. A., Smith, A. K., Flack, T. R., Gouws, A. D., & Andrews, T. J. (2021). Power contours: Optimising sample size and precision in experimental psychology and human neuroscience. *Psychological Methods*, *26*(3), 295–314. https://doi.org/10/ghhqz7

Boudewyn, M. A., Luck, S. J., Farrens, J. L., & Kappenman, E. S. (2018). How many trials does it take to get a significant ERP effect? It depends. *Psychophysiology*, *55*(6), e13049. https://doi.org/10/gcr25d

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10/k9d

Clayson, P. E., Baldwin, S. A., Rocha, H. A., & Larson, M. J. (2021). The data-processing multiverse of event-related potentials (ERPs): A roadmap for the optimization and standardization of ERP processing and reduction pipelines. *NeuroImage*, *245*, 118712. https://doi.org/10/gnn9gr

Clayson, P. E., Carbine, K. A., Baldwin, S. A., & Larson, M. J. (2019). Methodological reporting

behavior, sample sizes, and statistical power in studies of event-related potentials: Barriers to reproducibility and replicability. *Psychophysiology*, *56*(11), e13437. https://doi.org/10/ggnvxz

Cohen, J. (1988). *Statistical power analysis for tibe behavioral sciences*. Hillsdale, NJ, l, awrence Erlbanin.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

Cohen, J., & Polich, J. (1997). On the number of trials needed for P300. *International Journal of Psychophysiology*, *25*(3), 249–255. https://doi.org/10/c5zjsw

Delorme, A. (2023). EEG is better left alone. *Scientific Reports*, *13*(1), 2372. https://doi.org/10/gsj78h

Duncan, C. C., Barry, R. J., Connolly, J. F., Fischer, C., Michie, P. T., Näätänen, R., Polich, J., Reinvang, I., & Van Petten, C. (2009). Event-related potentials in clinical research: Guidelines for eliciting, recording, and quantifying mismatch negativity, P300, and N400. *Clinical Neurophysiology*, *120*(11), 1883–1908. https://doi.org/10/d3gz5h

Fischer, A. G., Klein, T. A., & Ullsperger, M. (2017). Comparing the error-related negativity across groups: The impact of error-and trial-number differences. *Psychophysiology*, *54*(7), 998–1009. https://doi.org/10/gspvpq

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PloS One*, *9*(10), e109019. https://doi.org/10/gfpf3m

Gibney, K. D., Kypriotakis, G., Cinciripini, P. M., Robinson, J. D., Minnix, J. A., & Versace, F. (2020). Estimating statistical power for event-related potential studies using the late positive potential. *Psychophysiology*, *57*(2). https://doi.org/10/gmbkzv

Hall, L., Dawel, A., Greenwood, L., Monaghan, C., Berryman, K., & Jack, B. N. (2023). Estimating statistical power for ERP studies using the auditory N1, Tb, and P2 components. *Psychophysiology*, e14363. https://doi.org/10.1111/psyp.14363

Huffmeijer, R., Bakermans-Kranenburg, M. J., Alink, L. R., & Van IJzendoorn, M. H. (2014). Reliability of event-related potentials: The influence of number of trials and electrodes. *Physiology & Behavior*, *130*, 13–22. https://doi.org/10/f55v9v

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124.

Jensen, K. M., & MacDonald, J. A. (2023). Towards thoughtful planning of ERP studies: How participants, trials, and effect magnitude interact to influence statistical power across seven ERP components. *Psychophysiology*, *60*(7), e14245. https://doi.org/10/gr8z6f

Kappenman, E. S., & Luck, S. J. (2010). The effects of electrode impedance on data quality and statistical significance in ERP recordings. *Psychophysiology*, *47*(5), 888–904.

Kiesel, A., Miller, J., Jolicœur, P., & Brisson, B. (2008). Measurement of ERP latency differences: A comparison of single-participant and jackknife-based scoring methods. *Psychophysiology*, *45*(2), 250–274. https://doi.org/10/cckt4q

Larson, M. J., Baldwin, S. A., Good, D. A., & Fair, J. E. (2010). Temporal stability of the error-related negativity (ERN) and post-error positivity (Pe): The role of number of trials. *Psychophysiology*, *47*(6), 1167–1171.

Laszlo, S., Ruiz-Blondet, M., Khalifian, N., Chu, F., & Jin, Z. (2014). A direct comparison of active and passive amplification electrodes in the same amplifier system. *Journal of Neuroscience Methods*, *235*, 298–307. https://doi.org/10/f6h6hv

Luck, S. J., & Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology*, *54*(1), 146–157. https://doi.org/10/f9j228

Luck, S. J., & Kappenman, E. S. (2017). *Electroencephalography and event-related brain potentials.*

Marco-Pallares, J., Cucurell, D., Münte, T. F., Strien, N., & Rodriguez-Fornells, A. (2011). On the number of trials needed for a stable feedback-related negativity. *Psychophysiology*, *48*(6), 852–860. https://doi.org/10/brftk3

McQuitty, S. (2004). Statistical power and structural equation models in business research. *Journal of Business Research*, *57*(2), 175–183. https://doi.org/10.1016/S0148-2963(01)00301-0

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 1–9. https://doi.org/10/bw28

Ngiam, W. X. Q., Adam, K. C. S., Quirk, C., Vogel, E. K., & Awh, E. (2021). Estimating the statistical power to detect set-size effects in contralateral delay activity. *Psychophysiology*, *58*(5). https://doi.org/10/gr9cj5

Olvet, D. M., & Hajcak, G. (2009). The stability of error-related brain activity with increasing trials. *Psychophysiology*, *46*(5), 957–961. https://doi.org/10/cpf268

Picton, T. W. (2010). *Human auditory evoked potentials*. Plural Publishing.

Pontifex, M. B., Scudder, M. R., Brown, M. L., O'Leary, K. C., Wu, C.-T., Themanson, J. R., & Hillman, C. H. (2010). On the number of trials necessary for stabilization of error-related brain activity across the life span. *Psychophysiology*, *47*(4), 767–773.

Puce, A., & Hämäläinen, M. S. (2017). A review of issues related to data acquisition and analysis in EEG/MEG studies. *Brain Sciences*, *7*(6), 58.

Rietdijk, W. J., Franken, I. H., & Thurik, A. R. (2014). Internal consistency of event-related potentials associated with cognitive control: N2/P3 and ERN/Pe. *PloS One*, *9*(7), e102672. https://doi.org/10/f6hkpq

Sandre, A., Banica, I., Riesel, A., Flake, J., Klawohn, J., & Weinberg, A. (2020). Comparing the effects of different methodological decisions on the error-related negativity and its association with behaviour and gender. *International Journal of Psychophysiology*, *156*, 18–39. https://doi.org/10/gg55x4

Schweizer, G., & Furley, P. (2016). Reproducible research in sport and exercise psychology: The role of sample sizes. *Psychology of Sport and Exercise*, *23*, 114–122. https://doi.org/10/f8dm5n

Segalowitz, S. J., & Barnes, K. L. (1993). The reliability of ERP components in the auditory oddball paradigm. *Psychophysiology*, *30*(5), 451–459. https://doi.org/10/fc45g3

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384. https://doi.org/10/gc4jb7

Smulders, F. T. (2010). Simplifying jackknifing of ERPs and getting more out of it: Retrieving estimates of participants' latencies. *Psychophysiology*, *47*(2), 387–392. https://doi.org/10/fjttxk

Steele, V. R., Anderson, N. E., Claus, E. D., Bernat, E. M., Rao, V., Assaf, M., Pearlson, G. D., Calhoun, V. D., & Kiehl, K. A. (2016). Neuroimaging measures of error-processing: Extracting reliable signals from event-related potentials and functional magnetic resonance imaging. *Neuroimage*, *132*, 247–260. https://doi.org/10/f8jmrb

Thigpen, N. N., Kappenman, E. S., & Keil, A. (2017). Assessing the internal consistency of the event-related potential: An example analysis. *Psychophysiology*, *54*(1), 123–138. https://doi.org/10/f9j3cc

Ulrich, R., & Miller, J. (2001). Using the jackknife-based scoring method for measuring LRP onset effects in factorial designs. *Psychophysiology*, *38*(5), 816–827. https://doi.org/10/bndt8x

Zhang, B., & Stone, C. A. (2008). Evaluating item fit for multidimensional item response models. *Educational and Psychological Measurement*, *68*(2), 181–196. https://doi.org/10/fhq8ft

Statistical power analysis of event-related potential studies: methods and influencing factors

Nian Jingqing[1] Chen Xi[2] Chen Fangfang[3] Niu Xia[4] Luo Yu[1]

([1]School of Psychology, Guizhou Normal University, Guiyang, 550025, *China*)
([2]Shanghai OPPO Company, Shanghai, 200032, *China*)
([3] Wuhu Fourth People's Hospital, Wuhu, 241003, *China*)
([4] Anhui Medical University, Hefei, 230031, *China*)

Abstract

The robustness and reproducibility of research results are crucial to the development of scientific research, but complete statistical testing power reports are rarely seen in the event-related potential (ERP) research literature. This article mainly reviews and summarizes the existing research, thereby introducing statistical test power analysis methods, application examples, experimental design, effect amplitude, sample size, number of trials and other influencing factors in ERP research, with a view to providing researchers with design and / or pre-registration of research protocols and other stages that require calculation and reporting of statistical power in event-related potential studies to provide a reference.

Keywords: EEG; event-related potential; statistical power; sample size; number of trials